



UNIVERSITY OF
BIRMINGHAM



中华人民共和国农业部
Ministry of Agriculture,
P.R.China

Funded by Defra and the Ministry of Agriculture of the People's Republic of China

ECOGEOGRAPHIC DATA ANALYSIS: AN INTRODUCTION

Joana Magos Brehm¹, Nigel Maxted², Shelagh Kell²

¹Botanical Garden, National Museum of Natural History,
University of Lisbon (Portugal)

²School of Biosciences, University of Birmingham (UK)

CWR China

Second training workshop

Wednesday 12th January 2011

Contents



- Definition of ecogeographic study
- Ecogeographic survey model
- Data collection
- Data verification
- Data analysis

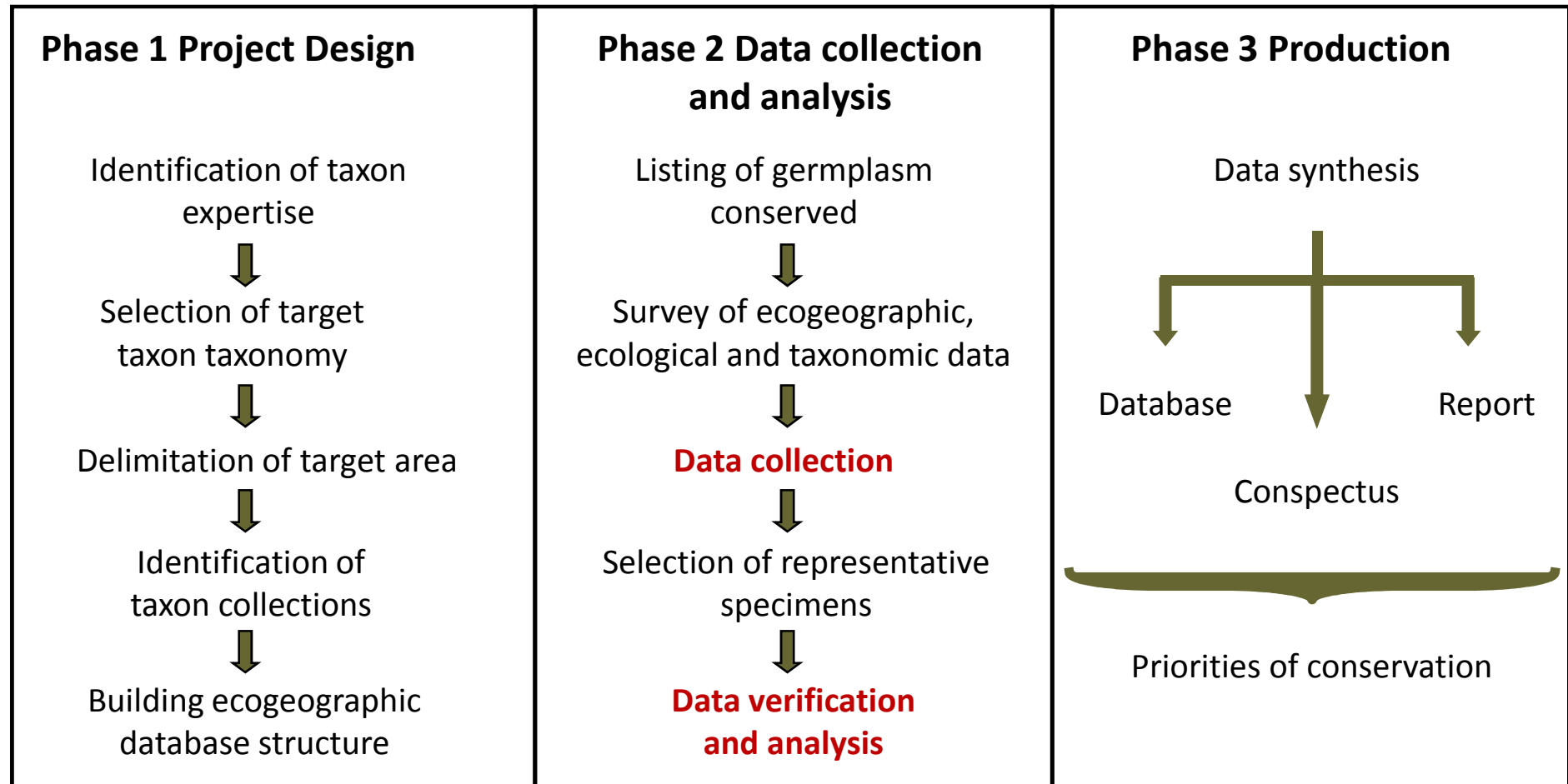
Ecogeographic study



"An **ecogeographic study** is an ecological, geographical and taxonomic information gathering and synthesis process. The results are predictive and can be used to assist in the formulation of collection and conservation priorities."

(Maxted *et al.* 1995)

Ecogeographic survey model



(adapted from Maxted *et al.* 1995)

Data collection



Four main categories of data types:

- Taxonomic
 - ▣ Taxon delineation
 - ▣ Degree of relatedness to crop taxon
 - ▣ Synonyms

- Distribution
 - ▣ Geographical taxon delineation
 - ▣ Taxon distribution

Data collection



Four main categories of data types (cont.):

- Conservation
 - Threat status
 - Legal data: conventions, policies
 - *In situ* and *ex situ*

- Environmental
 - Habitat, geological, topographical, climatic, soil and land use data
 - *Ex situ* characterization and evaluation data

Data collection

- Quality of data recorded
- Basic location but poor ecological data
- Hand written
- Foreign language
- Check identification



Data verification

- Assess completeness of the data set
 - ▣ certain analyses not possible if it is incomplete
- Check for errors (e.g. typing errors)
- Check for duplicates
- Assign coordinates (gazetteers, maps 1:25000, ENCARTA)
- Assign different levels of data accuracy
- Check for outlier locations (e.g. sea)



Data verification - example

Herbaria survey

- 10 Portuguese herbaria and 1 Spanish herbarium
- 3 online herbaria



(Magos Brehm 2009)

Data verification - example

Genebank survey

- 5 Portuguese genebanks
- 10 online genebanks

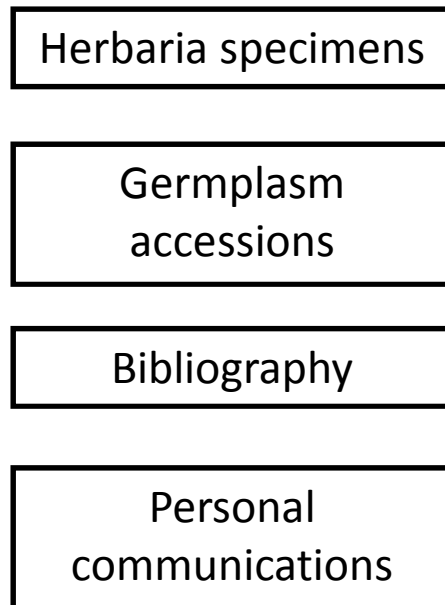


(Magos Brehm 2009)

Data verification - example

ACCESS Database

Sources of information



Field Name	Data Type	
ID_reg	Number	Unique identification number of the record
ID_SP	Text	Unique identification number of the species
ACR	Text	Acronym of the institution
Coll	Text	Particular collection
Data_sourc	Text	Source of data: H-herbarium, S-seed accession, R-bibliograph
Acc_Num	Text	Accession number
Gen_Name	Text	Name of the genus
Sp_Name	Text	Name of the species
Sp_Authors	Text	Authors of the species
Infra_Cat	Text	Subspecies or variety
Infra_Name	Text	Name of the infra category
Infra_Auth	Text	Authors of the infra category
Reg	Text	Portuguese administrative region
Loc	Text	Location
Level_inf	Text	Level of accuracy
Long	Number	Longitude
Lat	Number	Latitude
Gauss_X	Number	Gauss - X coordinate
Gauss_Y	Number	Gauss - Y coordinate
UTM_X	Number	UTM - X coordinate
UTM_Y	Number	UTM - Y coordinate
Long_dir	Text	Longitude direction
Lat_dir	Text	Latitude direction
Alt	Text	Altitude in metres
Date	Text	Collection date
Fl	Yes/No	If the specimen has flowers
Fr_imat	Yes/No	If the specimen has imature fruits
Fr_mat	Yes/No	If the specimen has mature fruits
No_fl_fr	Yes/No	If the specimen does not have either flowers or fruits
Ecol	Text	Ecological notes
Ass_Sp	Text	Associated species
Col	Text	Name of the collectors
Det	Text	Name of who identify the species
Obs	Text	Observations made by the collectors or identifiers
Rev1	Text	Revision notes
Rev2	Text	Revision notes
Rev3	Text	Revision notes
Rev4	Text	Revision notes
Rev5	Text	Revision notes
Photo	Yes/No	If there is photo
Photo_link	Hyperlink	Link to the photo
Source	Text	The source of information when not an herbarium specimen

(Magos Brehm 2009)

Data verification - example

Type of information obtained

HERBARIA SPECIMENS DATA

- ❖ Acronym
- ❖ Accession number
- ❖ Genus
- ❖ Species
- ❖ Infra

GERMPLASM ACCESSIONS DATA

- ❖ Authors
- ❖ Location
- ❖ Latitude
- ❖ Longitude
- ❖ Altitude
- ❖ Habitat
- ❖ Collectors
- ❖ Identify by
- ❖ Ecological data (type of soil, parent rock,...)

Data verification - example

Assign different levels of data accuracy

LEVEL OF ACCURACY	TYPE OF DATA
1	Municipalities (“concelhos”), between cities, mountains, big rivers
2	Cities, between parishes, around villages up to 15 Km
3	Parishes (“freguesias”), between settlements within the same parish, small mountains
4	Settlements, beaches, small streams, lagoons, capes, geodesic landmarks
5	Exact places, fountains, farms
?	Unknown location

Data verification - example

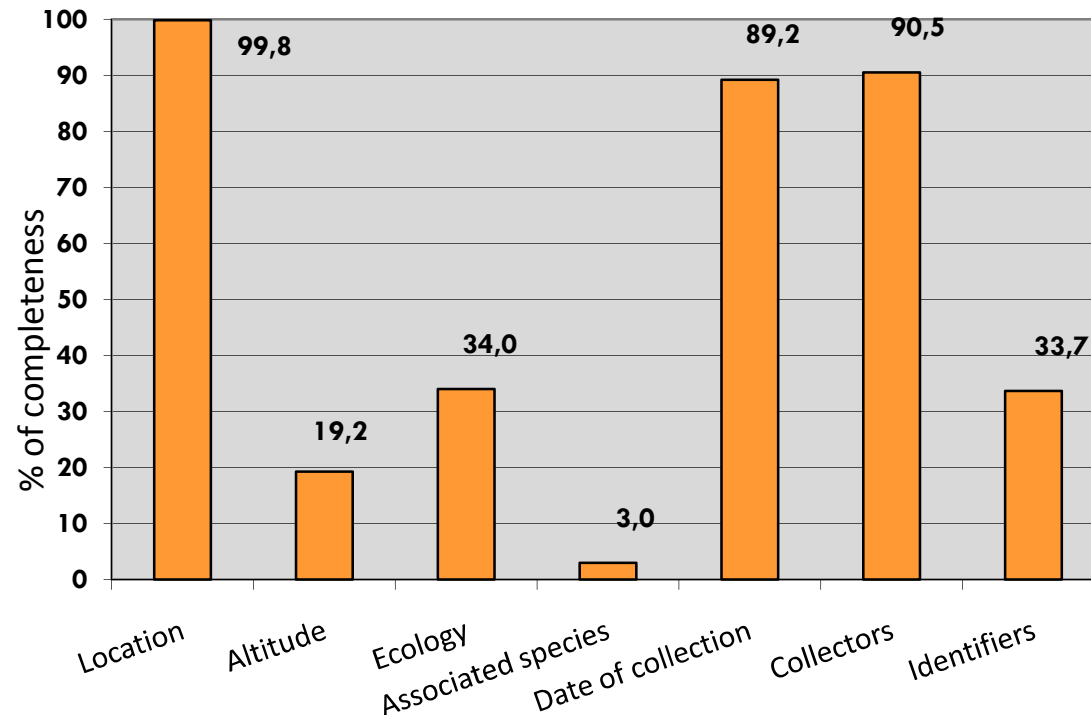
Herbaria specimens
483 records

Germplasm accessions
2 records

Bibliography
109 records

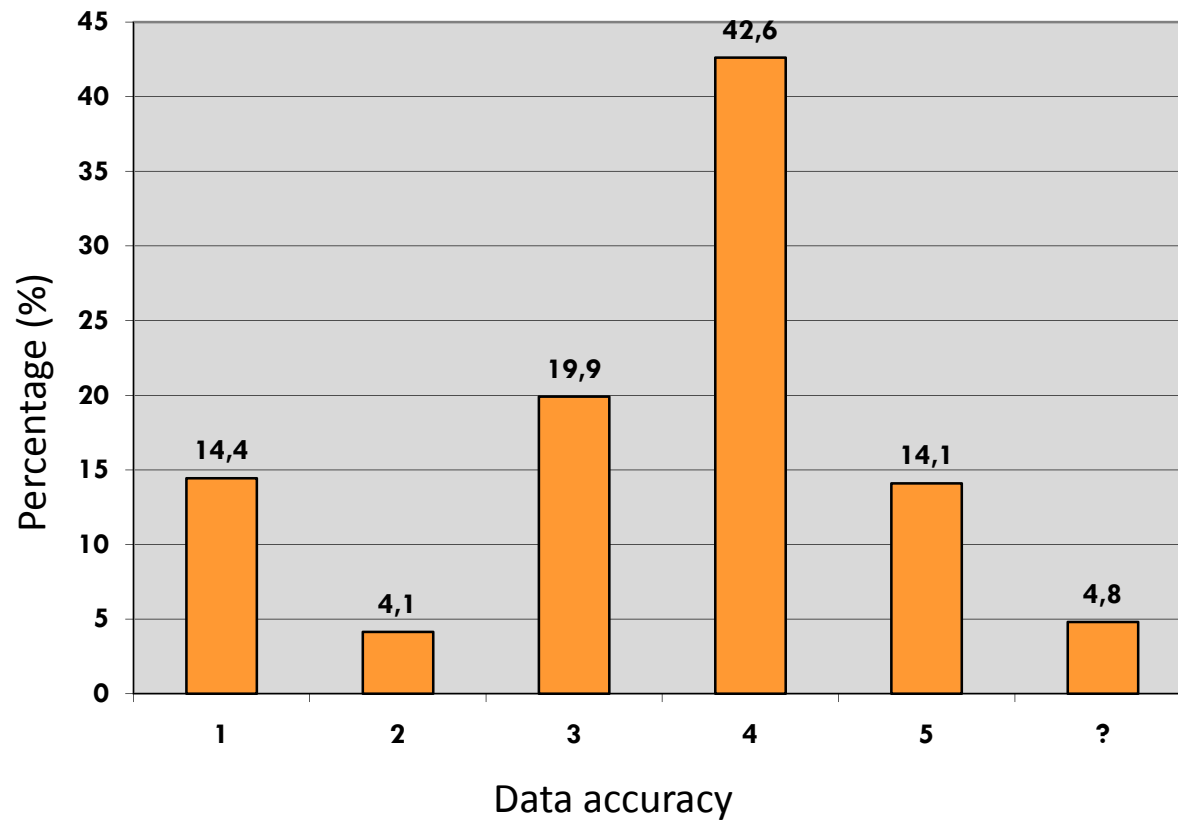
Personal communications
9 records

Total 603



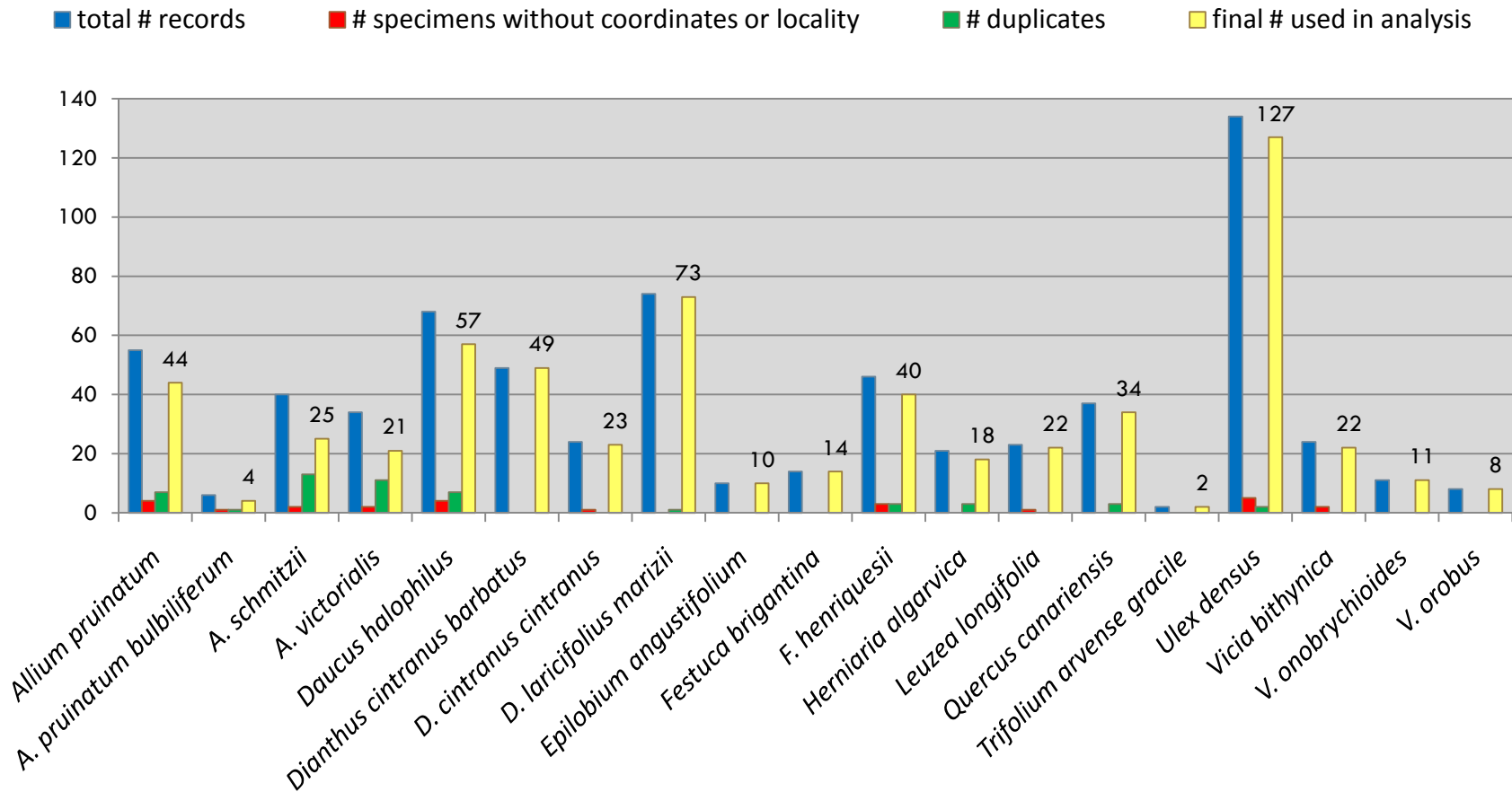
(Magos Brehm 2009)

Data verification - example



(Magos Brehm 2009)

Data verification - example



(Magos Brehm 2009)

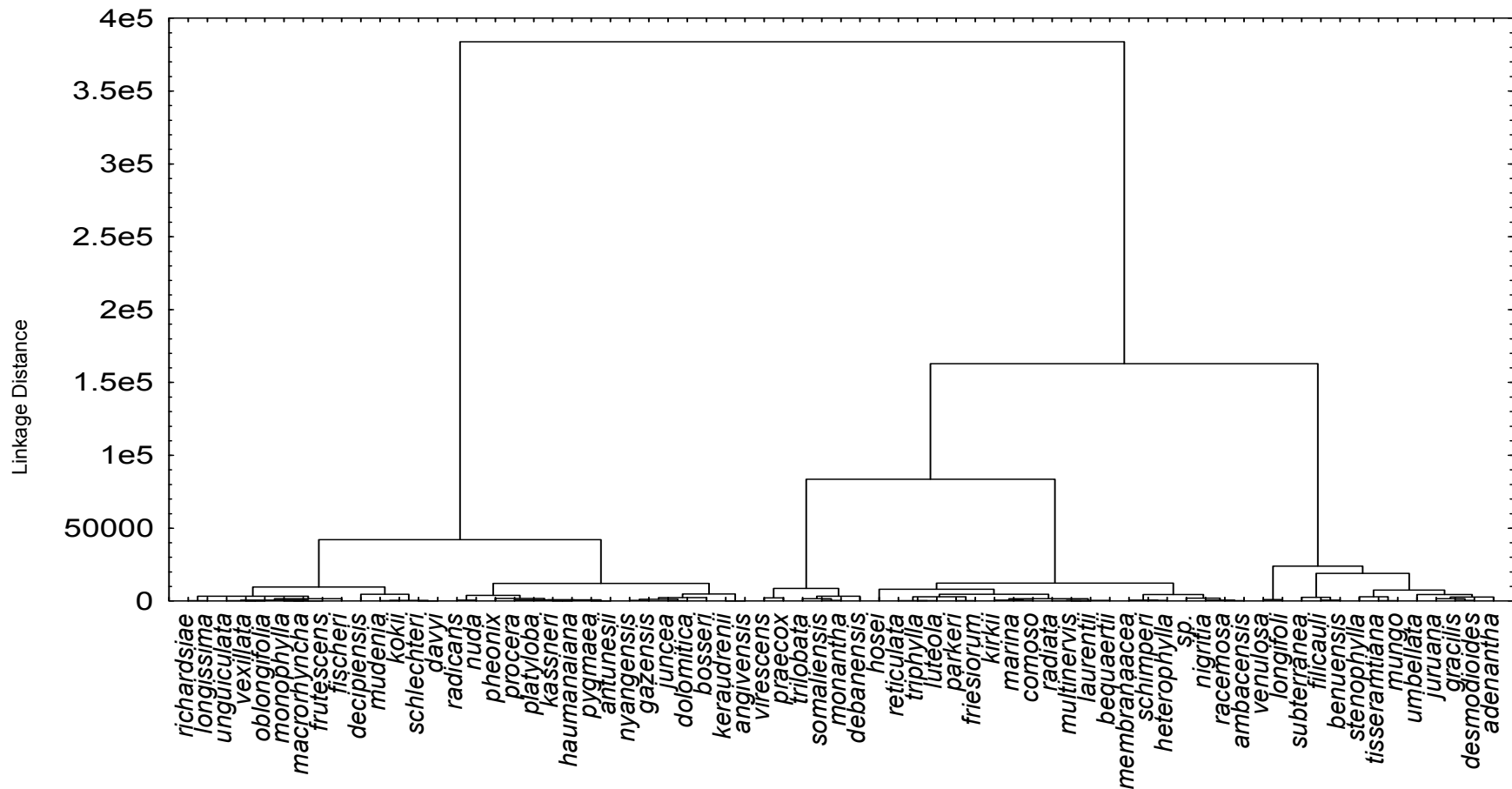
Data analysis



- Tools
 - ▣ Multivariate analysis (analyse more than one variable at a time)
 - cluster analysis
 - ordination techniques (PCA, PCoA, CVA)
 - ▣ Geographic information systems (GIS)
 - maps

Data analysis – multivariate analysis

Dendrogram of results from the multivariate cluster analysis (Ward's method) for climate variables per species



Data analysis



Ordination techniques

- Principal components analysis (PCA)
 - ▣ produces new quantitative variables which reflect the variation in the original descriptors

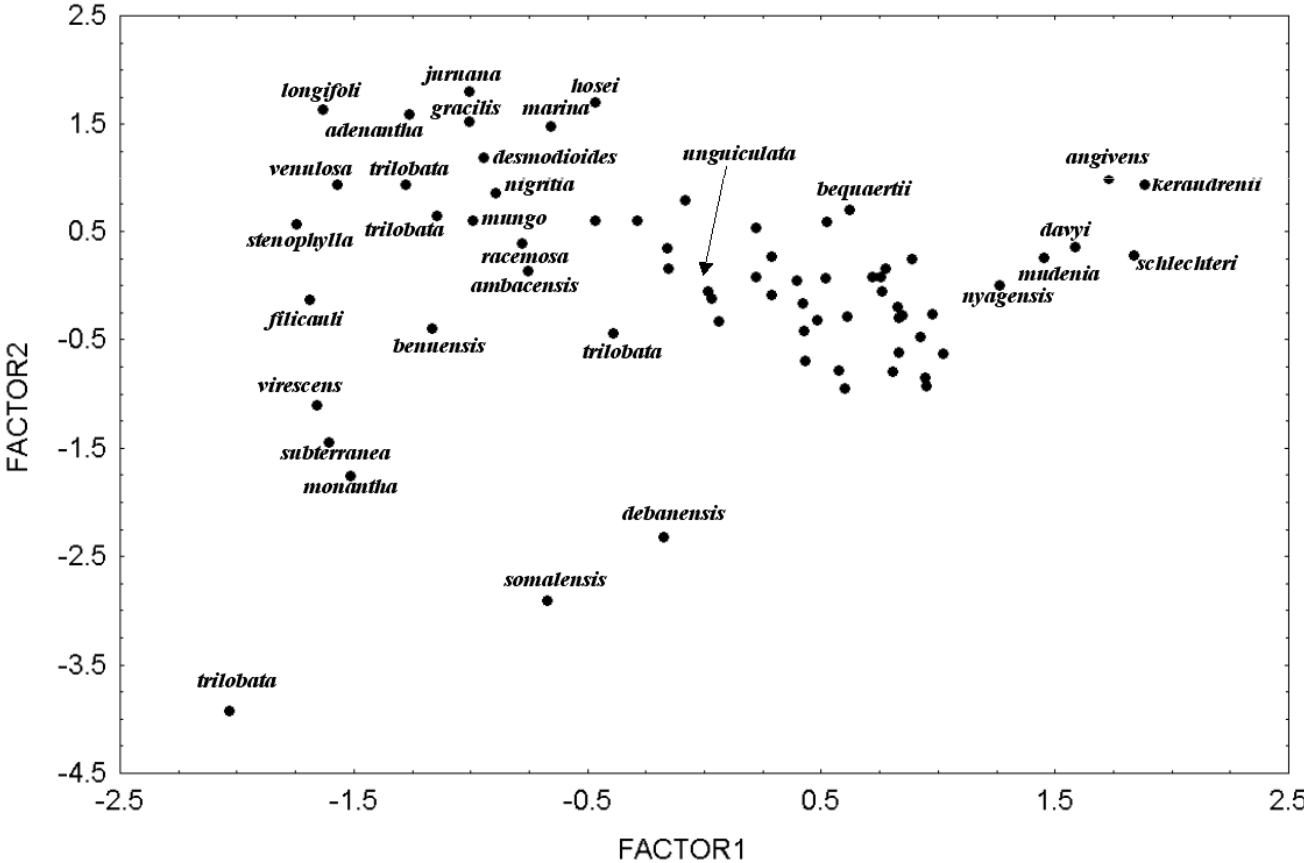
- Principal coordinates analysis (PCoA)
 - ▣ similar to PCA, for quantitative and qualitative data

- Canonical variate analysis (CVA)
 - ▣ distinctiveness of pre-defined groups

Data analysis– multivariate analysis

Principal components plot of climatic adaptations for each species

PCA1 = temperature & PCA2 = rainfall



Data analysis - GIS



Geographic Information Systems

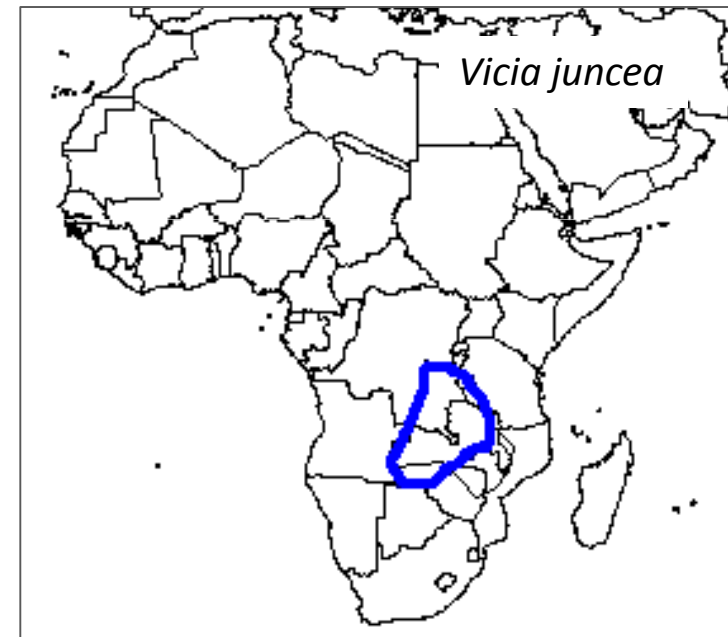
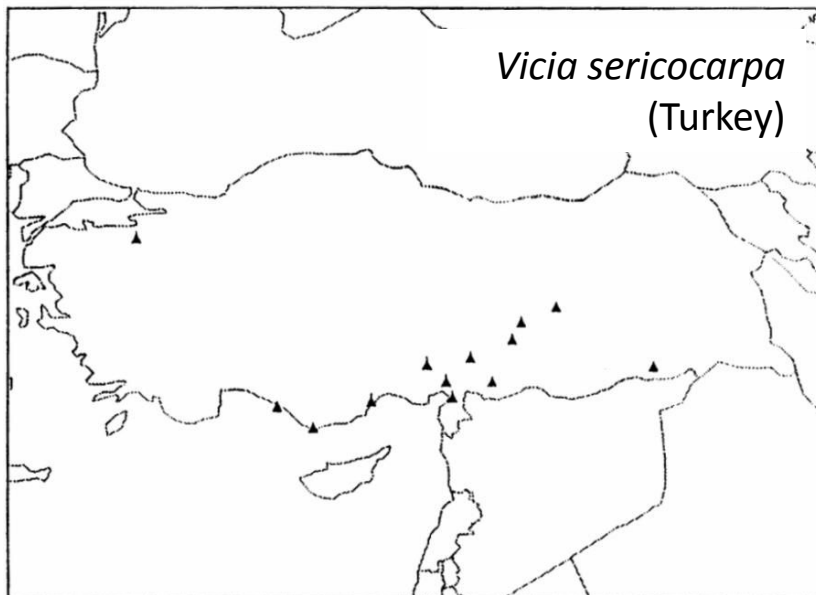
“GIS are integrated systems of computer hardware and software for the analysis and display of spatially distributed data”

(Johnston 1998)

(after coffee break!)

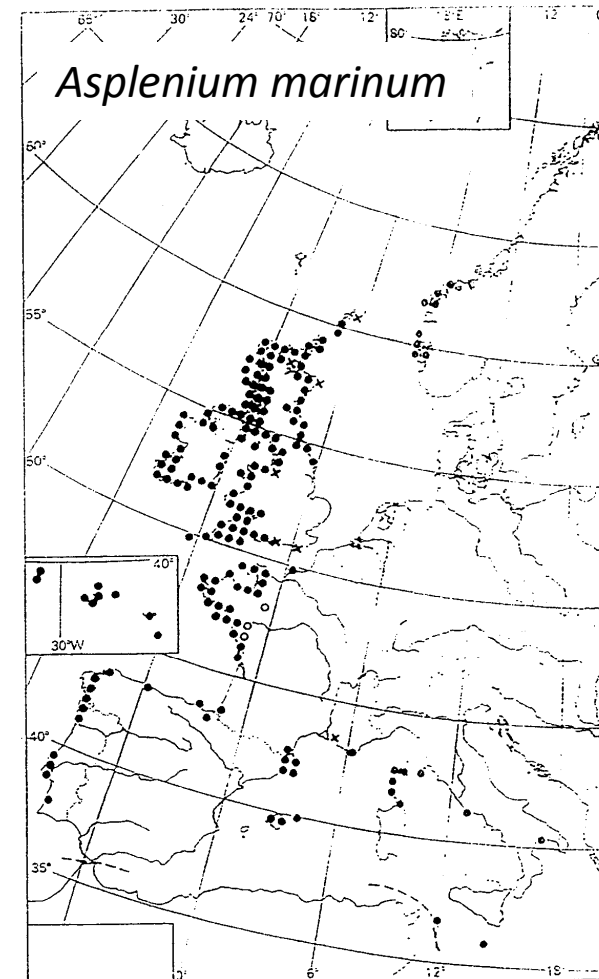
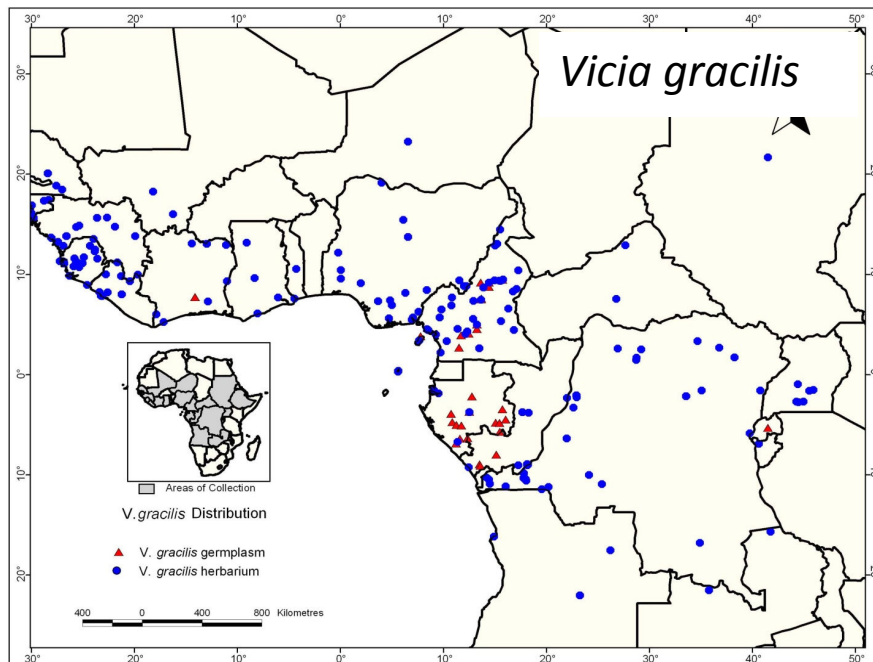
Data analysis - maps

- Distribution maps
 - Enclosed line
 - no indication of frequency
 - no local variation
 - problems with isolated occurrences



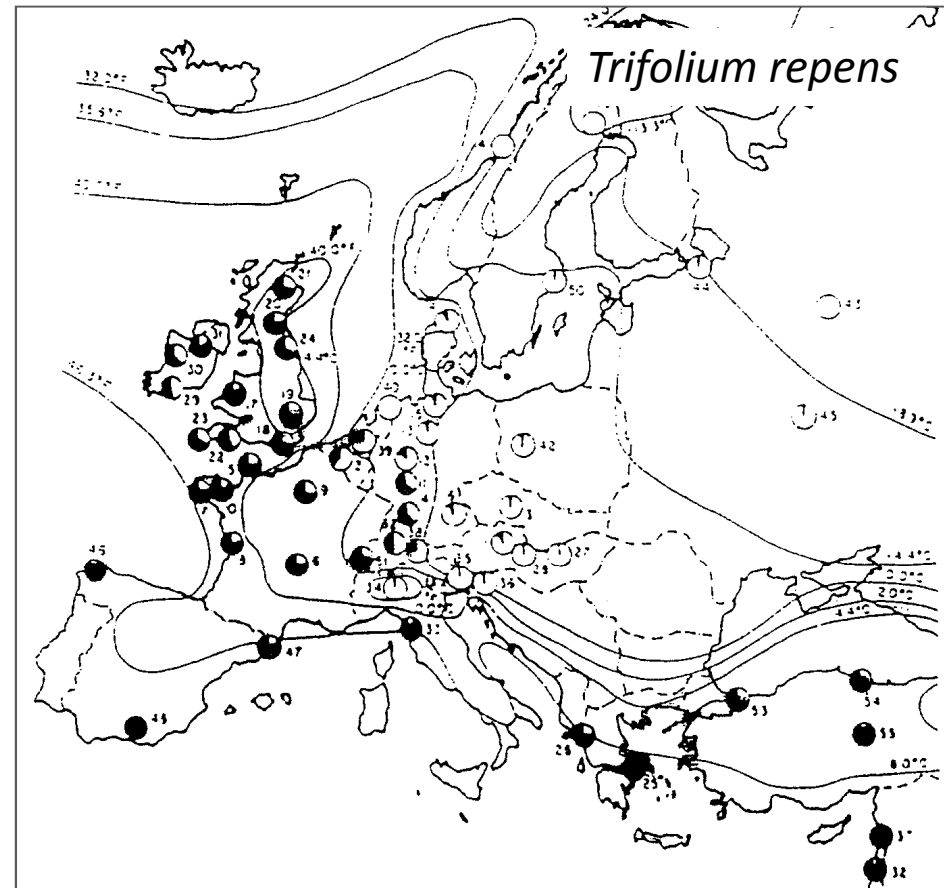
Data analysis - maps

- Distribution maps
 - Enclosed line
 - Dot distribution (point, grid, enhanced)



Data analysis - maps

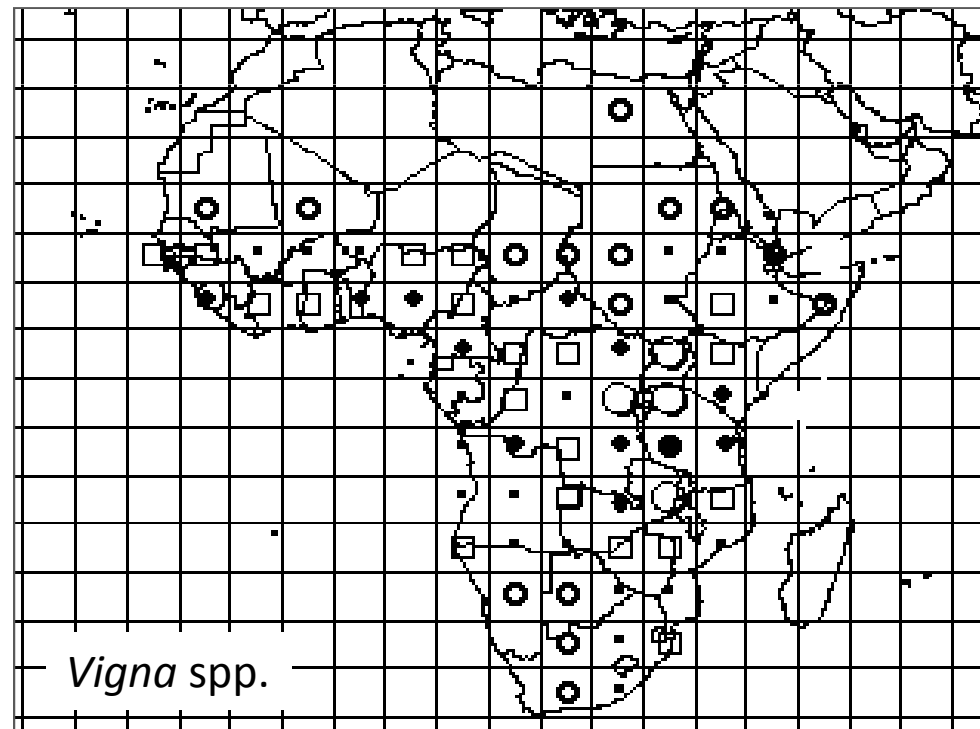
- Distribution maps
 - Enclosed line
 - Dot distribution (point, grid)
 - Enhanced dot distribution



Data analysis - maps

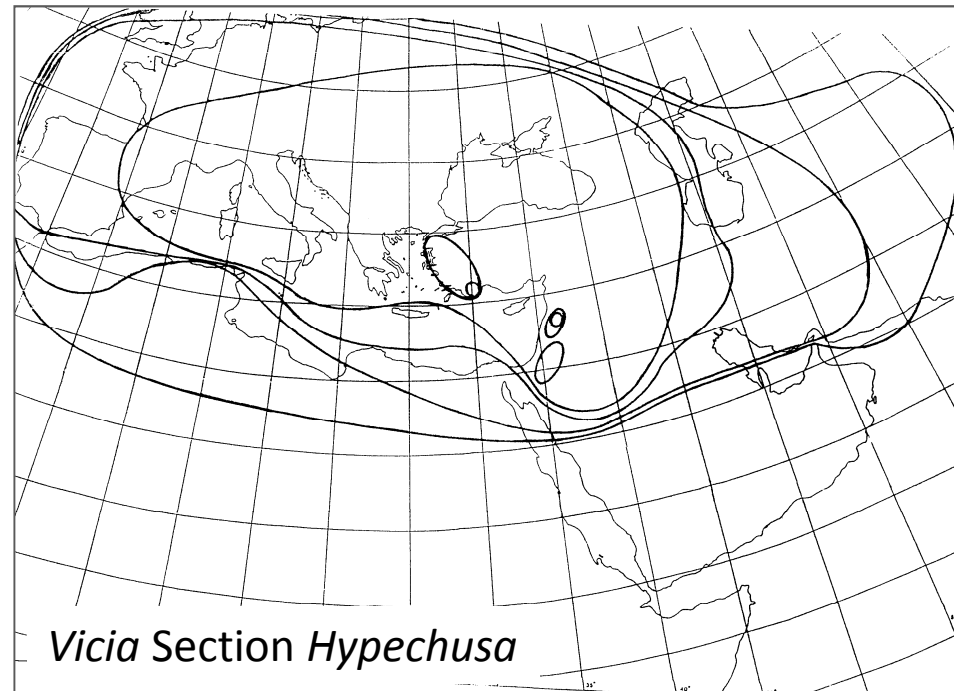
- Distribution maps
 - Enclosed line
 - Dot distribution (point, grid)
 - Enhanced dot distribution

- 1-5 species
- 6-10 species
- 11-15 species
- ◆ 16-20 species
- 21-25 species
- ◆ 26-30 species
- 31+ species



Data analysis - maps

- Distribution maps
 - Enclosed line
 - Dot distribution (point, grid)
 - Enhanced dot distribution
 - Contour map



Data analysis



- Distribution maps
- Ecogeographic characterisation of populations/species (to help interpret geographic, ecological and taxonomic patterns)

Display relative frequencies of specimens/characters from different sites and with environmental variables

Data analysis – ecogeographic characterisation



Ecogeography studies the **effect of environment** on living organisms

ECOLOGICAL DESCRIPTORS

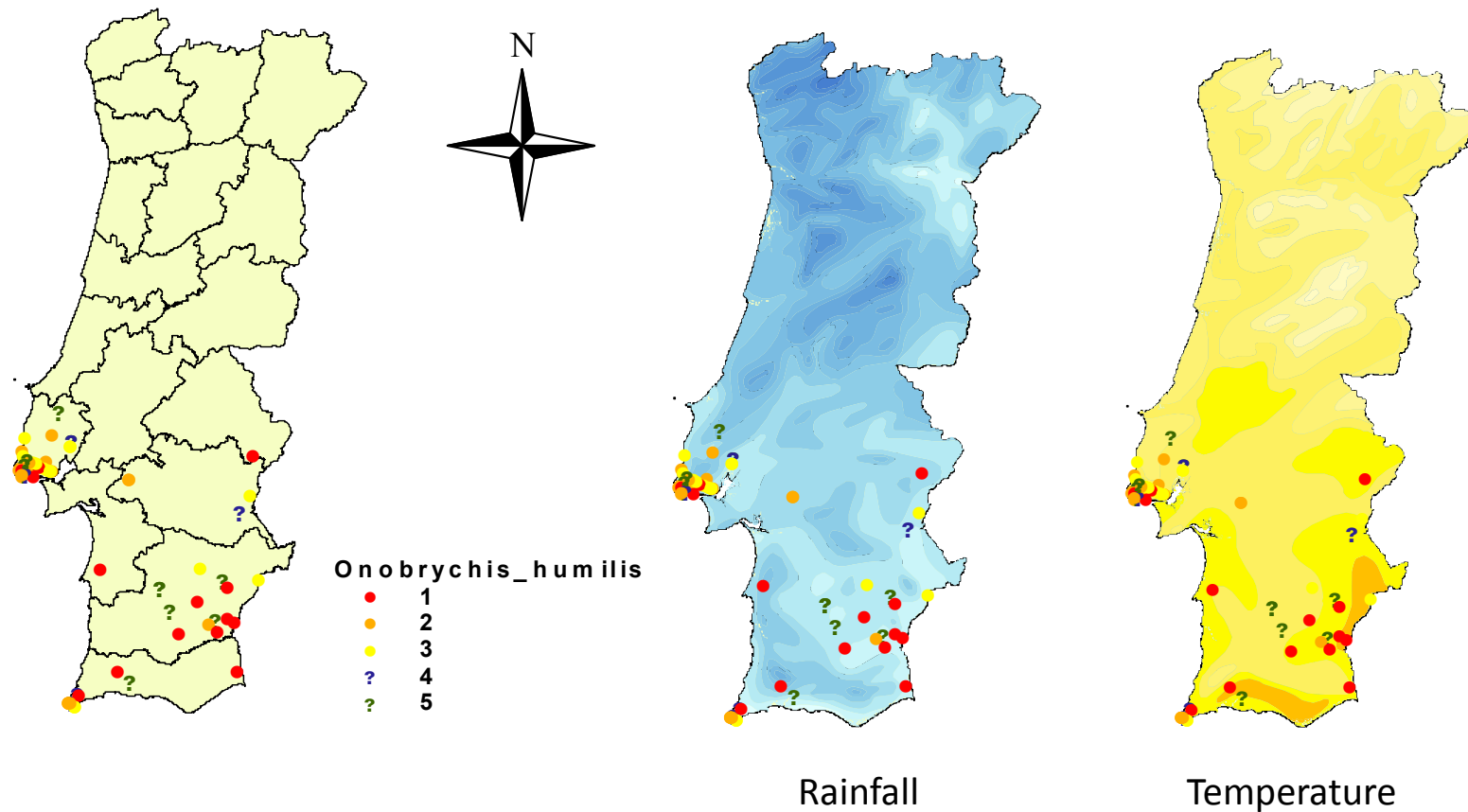
- Biotic factors: other species (competitors, pollinators, etc.)

- Abiotic factors:
 - Temperature, precipitation, climate indexes, etc.
 - Slope, aspect, hill shade, topographic shape, etc.
 - Soil type, pH, salinity, organic carbon, etc.

- Anthropogenic factors

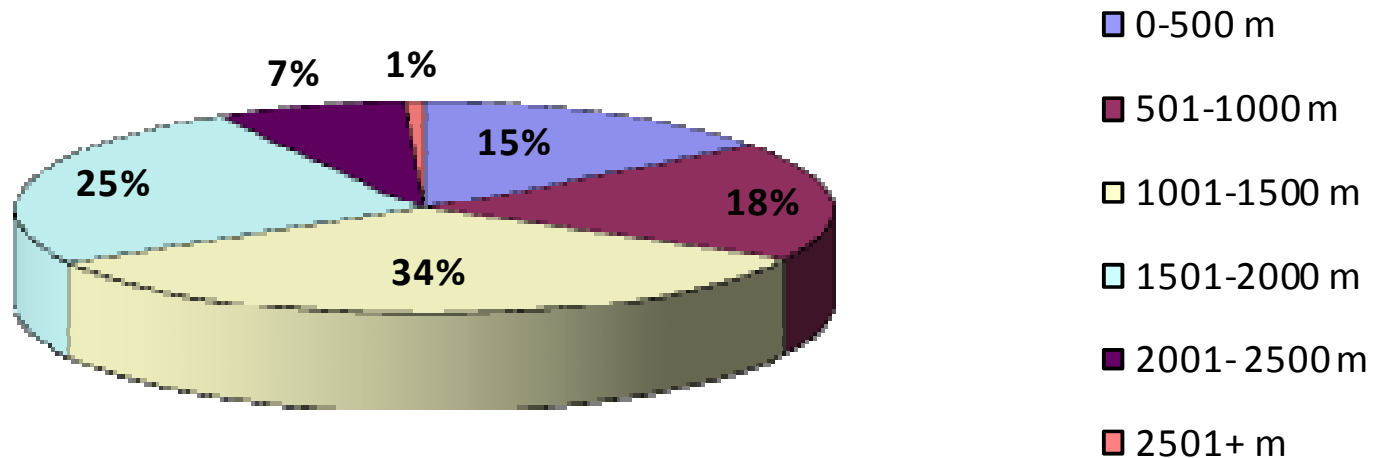
Data analysis – ecogeographic characterisation

Onobrychis humilis (L.) G. López



Data analysis – ecogeographic characterisation

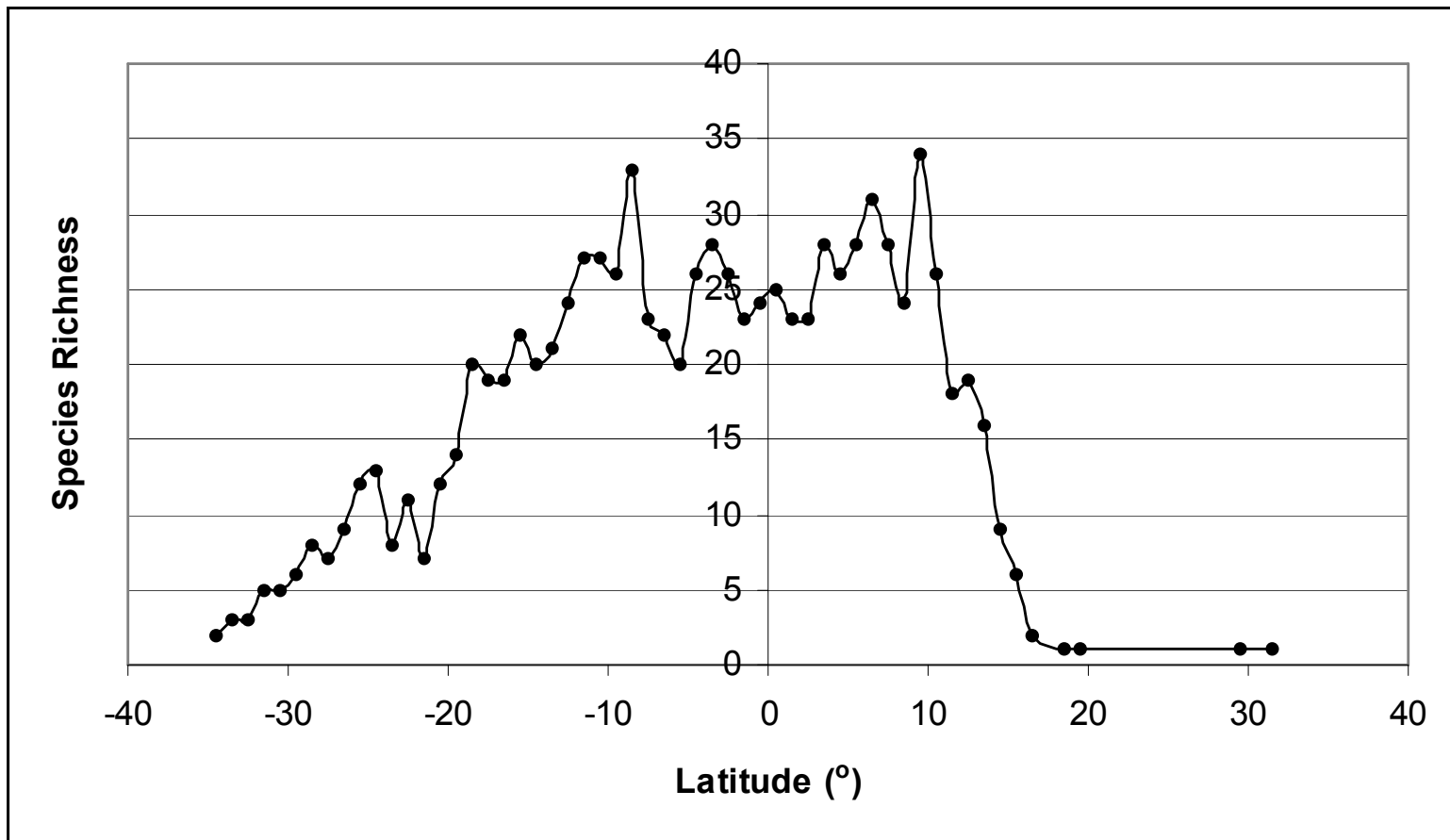
Proportion of all *Vigna* specimens collected at various altitudes in Africa



(Maxted *et al.* 2004)

Data analysis – ecogeographic characterisation

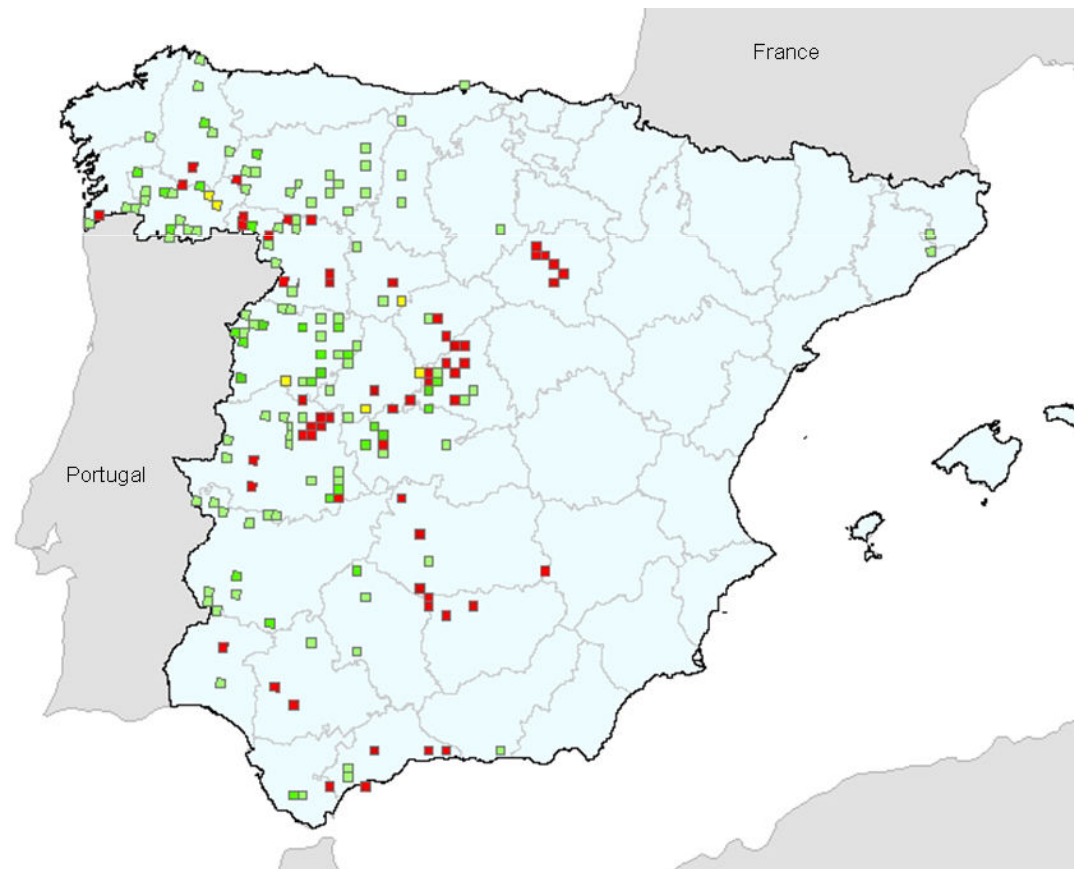
Number of *Vigna* species plotted against latitude



(Maxted *et al.* 2004)

Data analysis – ecogeographic characterisation

Lupinus luteus in Spain - select good quality georeferenced presence data



(Parra-Quijano *et al.* 2008)

Data analysis – ecogeographic characterisation



Lupinus luteus in Spain (cont.)

- Compile ecogeographical layers/variables (passport data, georeferencing collecting sites)
- Select ecogeographic variables
 - Subjective (experts knowledge)
 - Objective (statistics significance)
- A PCA can be undertaken in order to reduce the number of variables
- Create tables with accessions and the corresponding ecological descriptors

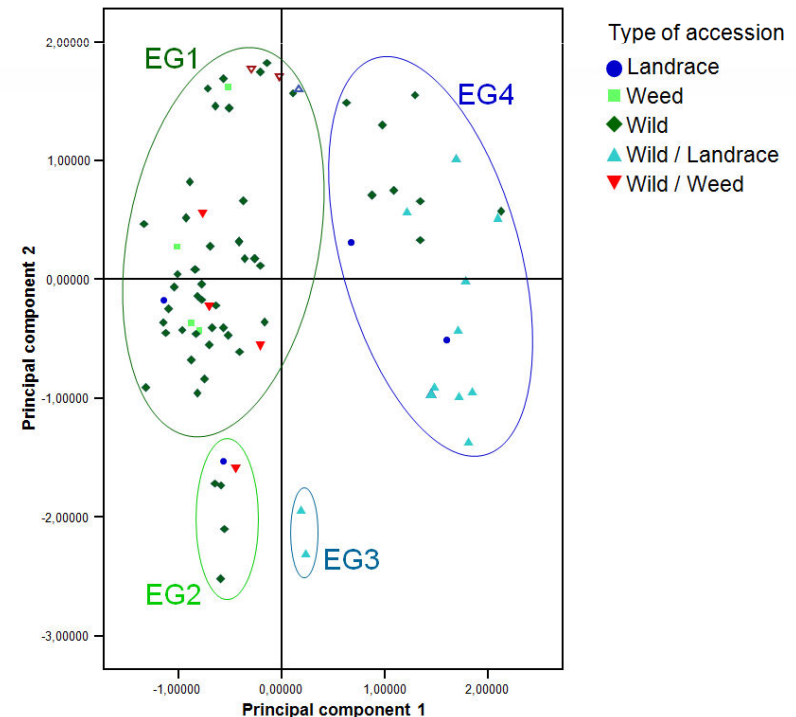
Data analysis – ecogeographic characterisation

Lupinus luteus in Spain (cont.)

- A PCA can be undertaken in order to reduce the number of variables
- Create tables with accessions and the corresponding ecological descriptors

Status of the accessions

Principal Component Analysis
3 components extracted
19 ecological descriptors



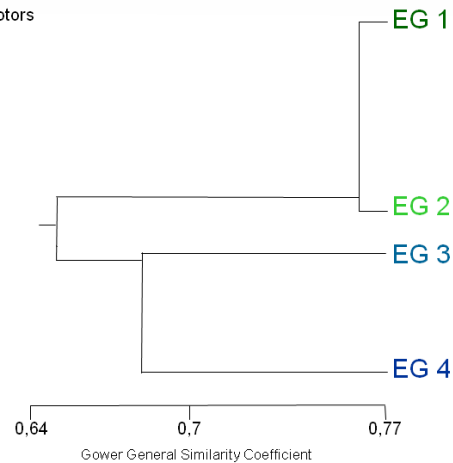
Data analysis – ecogeographical characterisation

Lupinus luteus in Spain (cont.)

- Estimate ecogeographical distances between all pairs of accessions (Gower similarity coefficient)
- Cluster analysis on the distance matrix and UPGMA agglomerative method (dendrograms that represent ecogeographical similarities between accessions)

L. luteus

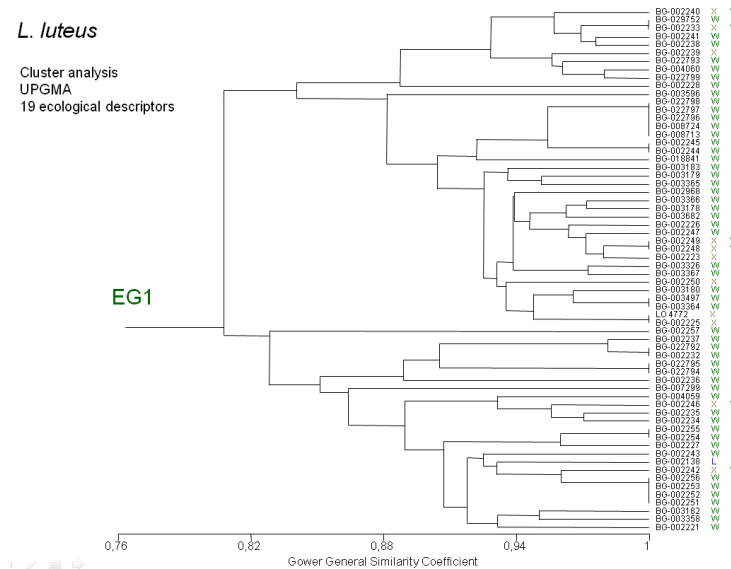
Cluster analysis
UPGMA
19 ecological descriptors



General ecogeographical grouping

L. luteus

Cluster analysis
UPGMA
19 ecological descriptors

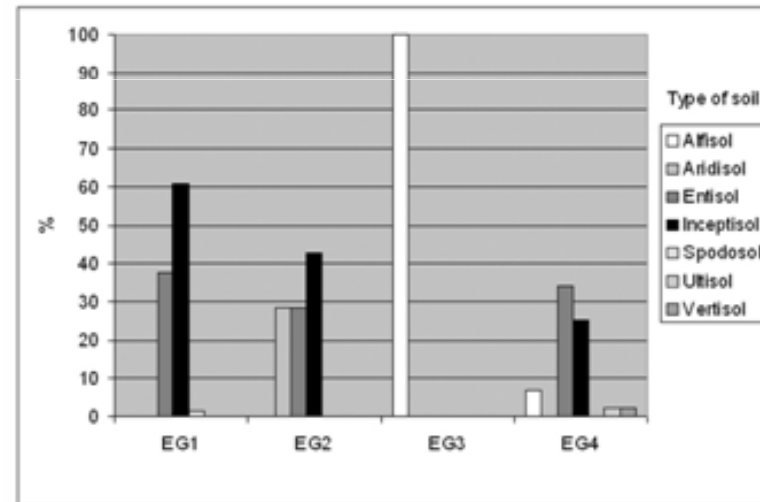
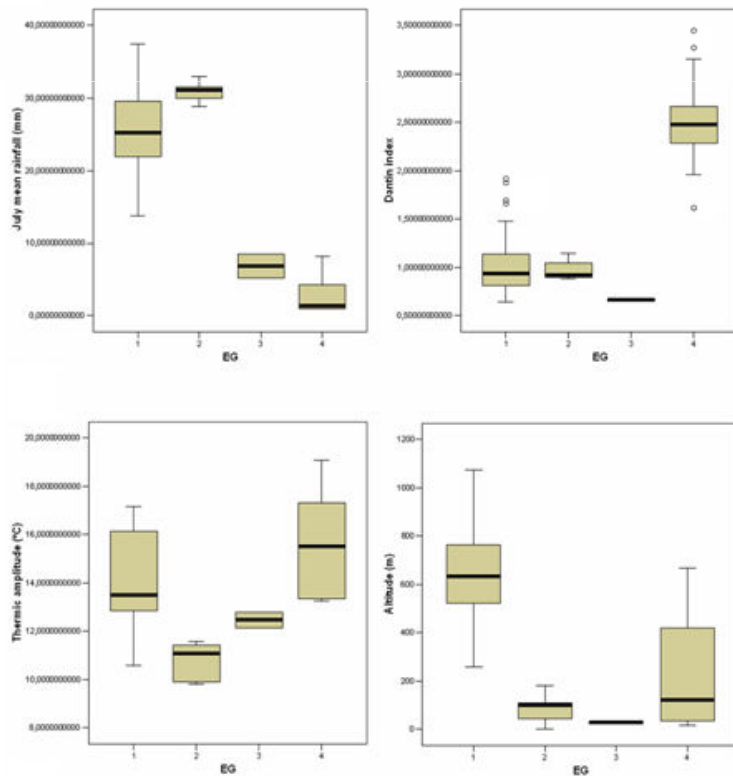


Detailed dendrogram for EG1 derived from the general dendrogram

Data analysis – ecogeographic characterisation

Lupinus luteus in Spain (cont.)

- Characterise each EG from the cluster analysis using the new variables (PCA1 related to thermopluviometric factors, PCA2 related to temperature, PCA3 related to edaphic factors)



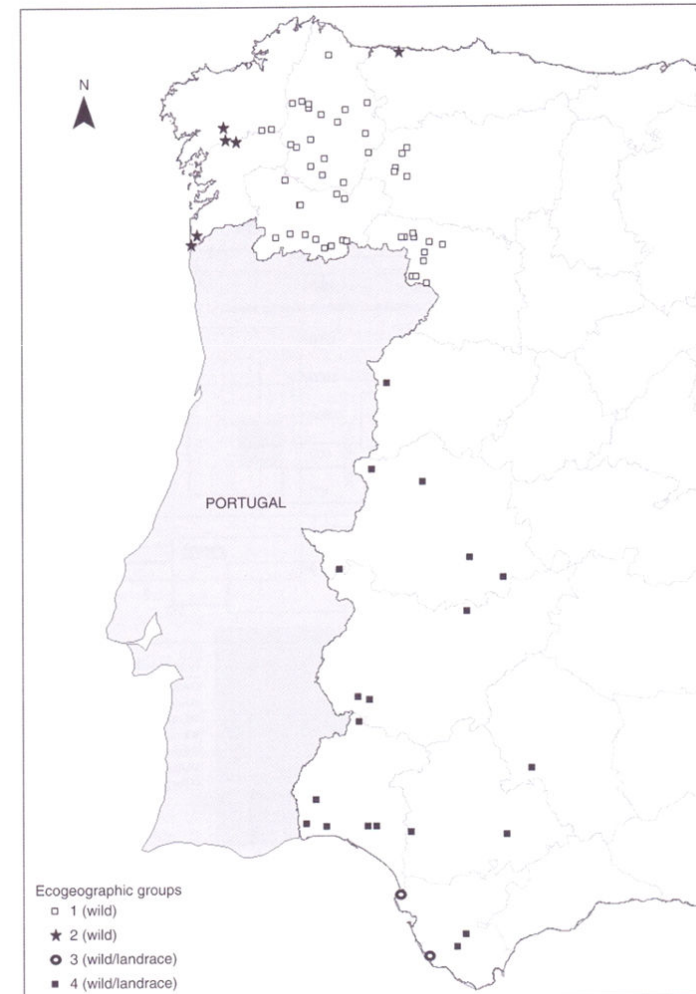
Differences between ecological groups (e.g. soil type)

(Parra-Quijano *et al.* 2008)

Data analysis – ecogeographic characterisation

Lupinus luteus in Spain (cont.)

- Assign each accession its corresponding EG and visualise them in a map



(Parra-Quijano *et al.* 2008)

Data analysis



- Distribution maps
- Ecogeographic characterisation of populations/species
- Ecogeographic representativeness (*in situ/ex situ*)

Data analysis – ecogeographic representativeness

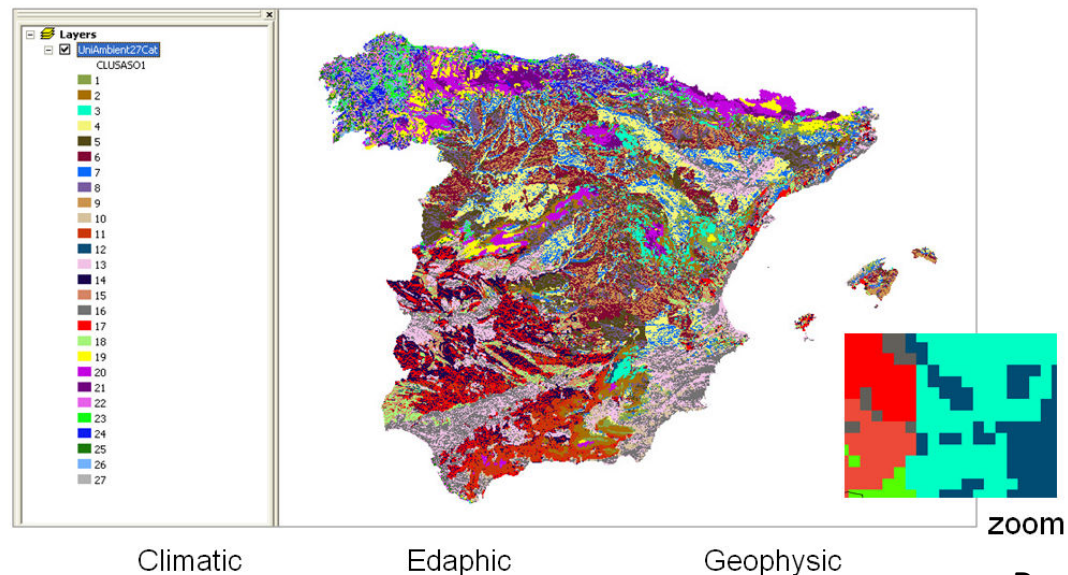


- Ecogeographical land characterisation - *Lupinus* CWR in Spain
 - ▣ Compile ecogeographical layers
 - ▣ Divide the spatial range into a grid
 - ▣ Extract environmental data for each of the grid
 - ▣ Cluster analysis in order to produce ecogeographical clusters (EC)
 - ▣ Descriptive statistics for each EC were obtained (quantitatively describe the environments represented)

Data analysis – ecogeographical representativeness

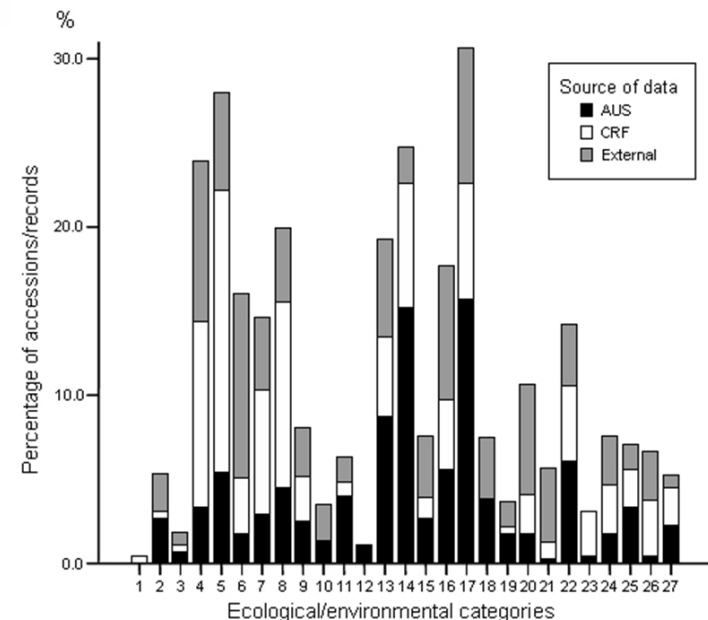
- Ecogeographical land characterisation - *Lupinus* CWR in Spain (cont.)
 - ▣ Each grid was assigned its corresponding ecogeographical category
 - ▣ The ecogeographical zones map is then created

Map of ecogeographical zones in Spain (27 environmental categories)



Data analysis – ecogeographic representativeness

- Ecogeographical land characterisation - *Lupinus* CWR in Spain (cont.)
 - Point maps were superimposed on the map of ecogeographical zones and each accession is assigned a value according to the ecogeographical zone where it was collected
 - The frequency and % of each of the ecogeographical zones represented in the accessions were obtained



Parra-Quijano *et al.* 2008; also see Parra-Quijano *et al.* 2011)

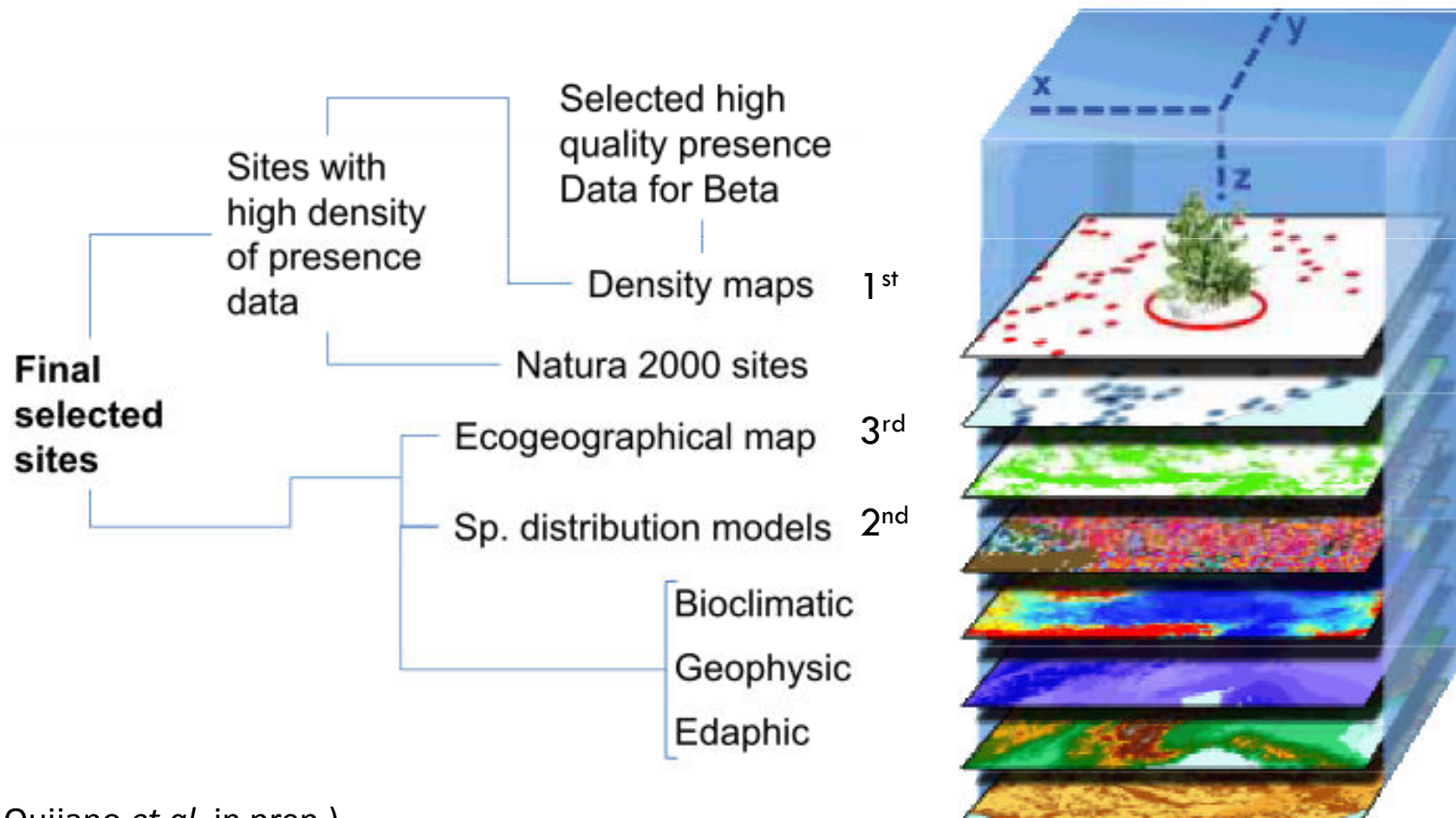
Data analysis



- Distribution maps
- Ecogeographic characterisation of populations/species
- Ecogeographic representativeness (*in situ/ex situ*)
- Genetic reserves set up

Data analysis – genetic reserves

- Genetic reserves set up - *Beta* CWR in Europe



(Parra-Quijano *et al.* in prep.)

Data analysis – genetic reserves

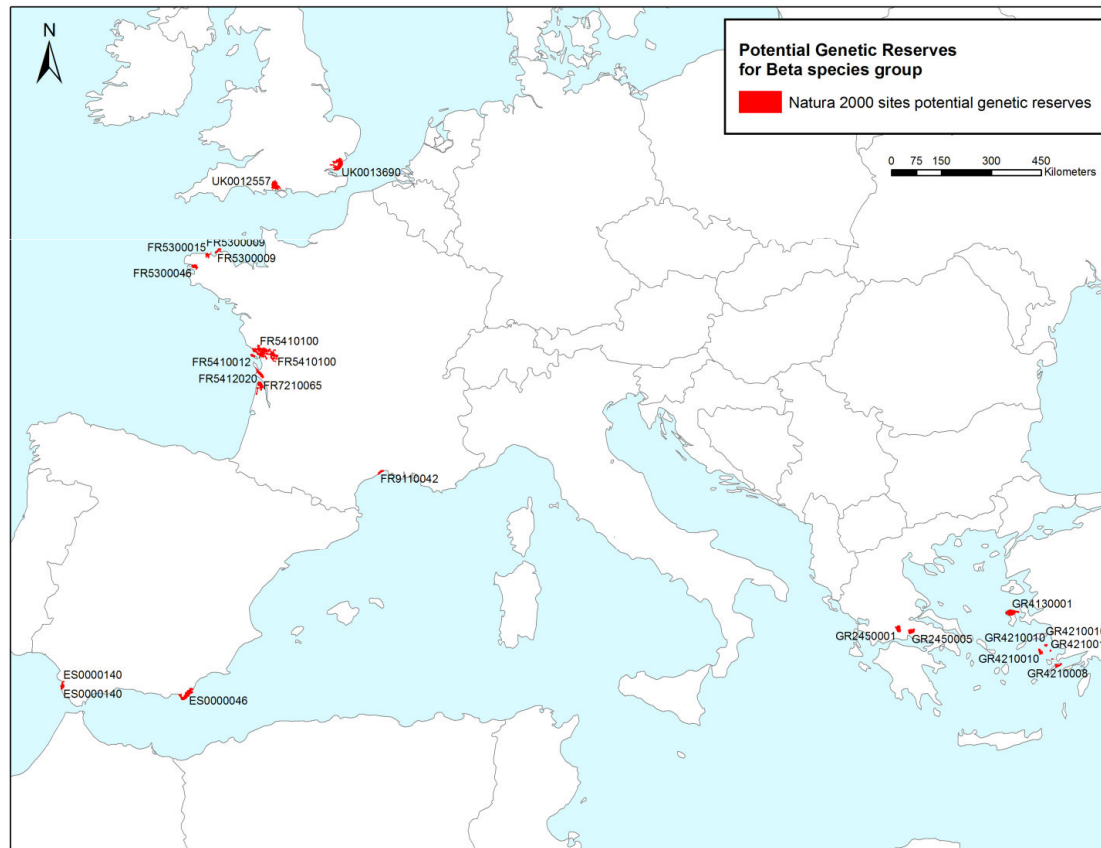
- Genetic reserves set up - *Beta* CWR in Europe (cont.)
 - safety criteria: sites are located within protected areas
 - efficiency criteria: they include areas with the greatest number of populations of the target species
 - representativeness criteria: they cover ecogeographical units where the species are commonly found as well as those where species are marginal.

Conservation of the species **greatest ecogeographical variability** is assumed to imply conservation of the **greatest genetic diversity** of adaptive importance and, possibly, the most interesting allelic variation in the genes of interest for plant breeding.

(Parra-Quijano *et al.* in prep.)

Data analysis – genetic reserves

- Genetic reserves set up - *Beta* CWR in Europe (cont.)



(Parra-Quijano *et al.* in prep.)



UNIVERSITY OF
BIRMINGHAM



中华人民共和国农业部
Ministry of Agriculture,
P.R.China

Funded by Defra and the Ministry of Agriculture of the People's Republic of China

ECOGEOGRAPHIC DATA ANALYSIS: AN INTRODUCTION

Joana Magos Brehm¹, Nigel Maxted², Shelagh Kell²

¹Botanical Garden, National Museum of Natural History,
University of Lisbon (Portugal)

²School of Biosciences, University of Birmingham (UK)

CWR China

Second training workshop

Wednesday 12th January 2011